On Testing of Hypotheses

Javier Rojo

Despite the fact that Erich L. Lehmann contributed substantially to several other areas of statistics, his name is prominently linked to hypothesis testing. This is not only because the tremendous impact of his 1959 book *Testing Statistical Hypotheses*, TSH, now in its third edition – Lehmann and Romano (2005), but it is also due to Erich's life-long pioneering and substantial contributions to the area. The chapter on Lehmann's books in the second volume of these selected works contains a fascinating account by Joseph P. Romano on the history of the writing of the substantially expanded third edition of *Testing Statistical Hypotheses*.

This chapter examines and summarizes some of Erich's early work on hypothesis testing. Much of this early work was incorporated into the books on hypothesis testing. The first edition of TSH contained three sections on sequential analysis, but these sections were not included in subsequent editions. In the preface to the second edition, Erich explains the decision to leave these sections out:

> *They are outdated and have been deleted, since it was not possible to do justice to the extensive and technically demanding expansion of this area. This is consistent with the decision not to include the theory of experimental design. Together with sequential analysis and survey sampling, this topic should be treated in a separate book.*

Lawrence D. Brown suggested their inclusion in this volume. Here's Lawrence's motivation:

> *The first edition of TSH has a beautiful treatment of the SPRT in Sections 3.10 - 3.12. This treatment was taken out of subsequent editions of TSH. (At the time, I tried to persuade Erich to leave this in, but he held his motivation for removing it, feeling that the theory was incomplete.) So far as I know no one else has written nearly as elegant a treatment since. It is still useful for contemporary students and academics to understand the fundamental ideas in this topic, even if the modern applications usually involve group sequential trials rather than ordinary one-by-one decisions, and composite hypotheses rather than only the simple vs simple special case treated here. It would thus be of interest - and useful - to reprint these 14 pages. I believe they can be read on their own, without reference to any earlier material in the text.*

J. Rojo
Professor of Statistics, Department of Statistics, Rice University, Houston, TX 77005
e-mail: jrojo@rice.edu

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_11, © Springer Science+Business Media, LLC 2012

Erich and I agreed. These three sections, sections 3.10 - 3.12, of TSH 1959, have been reproduced here.

The papers Lehmann (1947) and Lehmann and Stein (1948) deal with the same theme of testing a composite (null) hypothesis. The 1947 paper extends the work of Scheffé (1942). Let Θ denote a $k-$dimensional parameter space. Let Θ_0 denote the subset of Θ given by $\{\bar{\theta} \in \Theta : \theta_i = \theta_i^0\}$, for one $i = 1, \ldots, k$. Then the null hypothesis $H_0 : \bar{\theta} \in \Theta_0$ is an example of a composite (null) hypothesis with one constraint. Thus, in fact, the parameters $\theta_j, j \neq i$ are nuisance parameters. The case of more than one constraint is less fruitful, but results of Hsu (1945) are useful in this regard. Neyman (1935) had provided Type B regions for the case of a single nuisance parameter. These results were extended by Scheffé to the case of several nuisance parameters (under H_0), and Scheffé provided sufficient conditions for these Type B regions to also be Type B_1 (uniformly most powerful unbiased) regions. Of special interest, in Lehmann (1947), is the use of Neyman and Pearson's (1933) and Hsu's (1945) methods for finding the totality of similar regions as these regions play a central role in the existence of uniformly most powerful (UMP) or UMP one-sided regions in the case of composite hypotheses with one (or several) constraints. In Lehmann (1947), Erich extended Scheffé's results using the Neyman-Pearson method to obtain uniformly most powerful tests against one-sided alternatives and an example for testing for circular serial correlation in a normal population is presented; Hsu's method is also used to obtain UMP regions in cases, e.g. location & scale exponential and uniform distributions, where the method of Neyman and Pearson does not apply. This is achieved only after obtaining results that characterize the totality of similar regions for a large class of probability distributions that admit a sufficient statistic.

In Lehmann and Stein (1948) the problem of testing a composite hypothesis against a single alternative is addressed. Prior to this work, the approach to handle these problems utilized Neyman and Pearson's idea of restricting attention to similar regions. To fix ideas, let $\mathcal{F}$ denote a family of probability density functions and let g be a probability density function, $g \notin \mathcal{F}$. The test of interest is $H_0 : f \in \mathcal{F}$ against the alternative $H_1 : f = g$. For a level of significance α, Neyman and Pearson restricted attention to critical regions $\mathcal{W}$ such that $\int_{\mathcal{W}} f(x)dx = \alpha$ for all $f \in \mathcal{F}$. Such regions $\mathcal{W}$ are called *similar* regions. Hence a (similar) region is optimal of size α if it maximizes the power $\int_{\mathcal{W}} g(x)dx$ subject to $\mathcal{W}$ being *similar*. In this paper, Lehmann and Stein relaxed the condition of similarity to one requiring only that $\int_{\mathcal{W}} f(x)dx \leq \alpha$ or all $f \in \mathcal{F}$. By adapting the Neyman-Pearson lemma to apply in the present context, they derived sufficient conditions for most powerful tests. The newly developed theory is applied to some examples involving the normal family. Of special interest was Student's problem for which the composite null hypothesis consisted of the normal family with mean 0 and unspecified variance, while the simple alternative hypothesis consisted of the normal distribution with specified mean and variance. The result was somewhat surprising. Here is Erich's account as given in Lehmann (2008):

> *Of our four joint papers, I shall mention only one: "Most Powerful Tests of Composite Hypotheses" (1948). As the title indicates, the problem was to determine a level α test that would maximize the power against a specific alternative. Its purpose was to fill a gap in the classical Neyman-Pearson theory. This theory showed that many standard tests, for example*

the t-test, maximize the power among all unbiased tests. (A test is unbiased if its power against all alternatives is greater or equal to α.) But does this optimum property still hold when the restriction to unbiased tests is dropped and all level α tests are permitted to compete? We developed some general theory for this problem that naturally suggested itself as an adaptation from decision theory, and followed this by examining a number of classical testing problems. In some examples, the standard test retained its optimality against this wider competition; in others it did not. What was needed for these results was the construction of a "least favorable" weighted average of distributions in the hypothesis H as close to the alternative as possible. This least favorable distribution was often suggested by intuition, and then all went swimmingly. However, this turned out not to be the case when we came to the most interesting example, that of the t-test. We conjectured that the least favorable distribution would concentrate all its probability on a single point, but then intuition deserted us. We saw no way of determining this point and were quite frustrated. A few days later, Charles told me that he had solved the problem. Although not determining the point explicitly, he showed by a careful analysis that for $\alpha < 1/2$ such a point exists and gives the right test, which is quite different from the t-test. It has better power in the neighborhood of the specific alternative for which it was designed but lower power elsewhere. For $\alpha \geq 1/2$, the situation turned out to be much easier, and in that case the t-test cannot be improved.

The starting point for Lehmann (1950), Lehmann (1959), and Lehmann (2006) is the same – the likelihood ratio principle for testing. The three papers start by recognizing and lauding the intuition behind the likelihood ratio principle and the "reasonable" tests, in the sense that likelihood ratio tests usually agree with tests derived under optimality criteria, it produces. The three papers then proceed to examine different aspects of the testing problem motivated by the optimality of the likelihood ratio test in some cases, and its total failure in other cases.

In Lehmann (1959), Erich considers the case when the testing problem remains invariant under a group of transformations G. Furthermore, consider a class of invariant tests $\mathcal{F}$ endowed with an order that satisfies certain properties. Erich then proceeds to show that in this case, the optimality properties enjoyed by the likelihood ratio test follow directly from the fact that the test is uniformly most powerful invariant. The paper is also successful in unifying optimality results found in Kiefer (1958), who proved optimality results for symmetrical non-randomized designs, and results found in Wald (1942), who obtained optimality results for the analysis of variance test for the general univariate linear hypothesis. This unification of results under the same umbrella, is another clear example of Erich's system-building, rather than problem-solving, interests.

In Lehmann (2006) and Lehmann (1950), a closer examination of an ever-growing collection of examples where the likelihood ratio test becomes, for all practical purposes, useless, quickly changes the tone and each paper proceeds to examine, against this backdrop, the properties of tests produced by other approaches. When the testing problem remains invariant with respect to a transitive group of transformations, Lehmann (2006) proposed a testing approach – *the likelihood averaged or integrated with respect to an invariant measure approach*. In these cases the resulting test turns out to be uniformly at least as powerful as the corresponding likelihood ratio test, with the former being strictly better except when the two coincide. Moreover, even in the absence of the invariance property, the proposed approach continues to improve on the likelihood ratio test for many cases. In Lehmann (1950), Erich

embarked on a survey, and provided some extensions and modifications, of the existing theory. In particular, for example, the *Rao-Blackwellization* version of testing is provided. As a consequence, only tests which are functions of sufficient statistics need to be considered. When the problem remains invariant under certain groups of transformations, using the Hunt-Stein ideas, Erich discussed most stringent tests. Other approaches, such as unbiased tests, are also discussed.

The issue of "statistical significance" versus "practical significance" is addressed in Hodges and Lehmann (1954). Consider, for example, the problem of testing for the mean θ of a normal distribution. Suppose we consider the simple null hypothesis $H_0 : \theta = 0$ vs the alternative $H_a : \theta \neq 0$. Any reasonable test will have power increasing to 1 for every θ^* under the alternative, regardless of how close θ^* is to 0, as the sample size increases to ∞. Motivated by remarks in Berkson (1938), the paper proposes testing instead the null hypothesis that $\theta \in \Theta_0$, where Θ_0 is some neighborhood of 0 – an indifference zone –, and where the choice of Θ_0 is guided by the user's idea of practical significance for the problem at hand. The size of the test is then $\sup_{\theta \in \Theta_0} P_\theta(\textit{rejection})$. Several examples are discussed. Following is Erich's discussion of this paper as it appears in Lehmann (2008):

> *In the classical Neyman-Pearson theory, the hypothesis H:* $\theta = \theta_0$ *completely specifies the value of the parameter being tested. Frequently, it is more reasonable to consider instead the hypothesis H':* $|\theta - \theta_0| < \Delta$, *for some* $\Delta > 0$. *We worked out a number of examples, the most interesting (and rather complicated) being the case of a normal mean. The paper did not find much resonance, but the problem was later revived. It became known as "testing for bio-equivalence", however with the hypothesis and alternatives interchanged so that the hypothesis being tested was: H':* $|\theta - \theta_0| > \Delta$.

In the expository paper Lehmann (1993), Erich discusses the issue of the use of p-values versus the use of a predetermined significance level α, and the Fisher-Neyman feud over this issue. While acknowledging that there are other approaches to testing (Bayesian, fiducial) the paper focuses only on the $p-$value *vs.* predetermined significance level α debate. Erich's approach is, as it was in the case of the *data analytic vs. Bayesian vs. Frequentist* debate – see the chapter on Erich's philosophical work in this volume – to find common ground where both approaches complement, rather than contradict, each other. Erich had stated in various works and interviews that he did not think too much about foundational issues. See, for example, his comments included in the chapter on philosophical work. I commented there that his work demonstrated otherwise. Thus, it is interesting that in this paper he reaffirms this by stating that "I am not a philosopher, and this article is written from a statistical, not a philosophical, point of view". And while this is true, it is also true that the issue at hand is a foundational issue, although the topic might be considered of historical interest, an aspect that Erich was willing to accept. Thus in Lehmann (2008) he writes:

> *During the 1990's, I became interested in the history of statistics and found occasions to write about some aspects of Fisher's work. The first such instance was the surprising selection of me as the Fisher Lecturer for 1988 – surprising in view of the hostility between Neyman and Fisher and my close association with Neyman.*

And,

> *What has been emphasized by many authors is the difference between the calculation of p-values and the use of a predetermined significance level α, the former attributed to Fisher and the latter to Neyman and Pearson. However, in practice most users combine the two by calculating a p-value and then rejecting or accepting the hypothesis as the p-value is below or above α. This is in fact Fisher's own frequent practice. In 1993, I presented an account along these lines in an expository paper,* The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?

Erich concluded the paper by stating that an approach that takes into account, $p-$values, predetermined significance levels, power considerations, and conditioning, is feasible, while clearly giving conditioning a prominent place in the discussion. In connection with this last issue see the work of Kiefer (1975, 1976, 1977) and chapter 10 of TSH, 3^{rd} edition. This issue of conditioning also appears in an earlier paper – Lehmann (1958). In it, Erich discusses conditional tests in the exponential family setting, where conditioning is with respect to the complete sufficient statistic T. Noticing that, in some testing situations, the conditional power $\beta^*(t) = P_{\theta_1}\{rejecting \mid T = t\}$ at the alternative θ_1 depends on t, Erich then asks the question:

> *Suppose that $\beta^*(t)$ is quite small for the observed t, or quite high; is this value not more relevant to the case in hand than the average $\beta(\theta)$?*

Rather than answering the question directly, Erich then goes on to justify the use of $\beta^*(t)$ in a different way. Since T is complete and sufficient, then $\beta^*(T)$ becomes the best unbiased estimator for $\beta(\theta)$.

References

[1] J. Berkson. Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. *J. Am. Statist. Assoc.*, Vol. 33, No. 203, pp. 526-536, 1938.

[2] J. L. Hodges, Jr. and E. L. Lehmann. Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B*, Vol. 26, pp. 261-268, 1954.

[3] P. L. Hsu. On the Power Functions of the E^2-Test and the T^2-Test. *Ann. Math. Statist.*, Vol. 16, No. 3, pp. 278-286, 1945.

[4] J. Kiefer. Conditional confidence approach in multi-decision problems. In *Multivariate Analysis IV*, (P. R. Krishnaiah, ed.), pp. 143-158. Academic Press, New York, 1975.

[5] J. Kiefer. Admissibility of Conditional Confidence Procedures. *Ann. Statist.*, Vol. 4, No. 5, pp. 836-865, 1976.

[6] J. Kiefer. Conditional Confidence Statements and Confidence Estimators. *J. Am. Statist. Assoc.*, Vol. 72, No. 360, pp. 789-808, 1977.

[7] J. Neyman. Sur la vérification des hypothèses statistiques composées. *Bull. Soc. Math. France*, Vol. 63, pp. 1, 1935.

[8] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231, pp. 289-337, 1933.

[9] E. L. Lehmann. On optimum tests of composite hypotheses with one constraint. *Ann. Math. Statist.*, Vol. 18, pp. 473-495, 1947.

[10] E. L. Lehmann and C. Stein. Most powerful tests of composite hypotheses. I. Normal Distributions. *Ann. Math. Statist.*, Vol. 19, No. 4, pp. 495-516, 1948.

[11] E. L. Lehmann. Some principles of the theory of testing hypotheses. *Ann. Math. Statist.*, Vol. 21, pp. 1-26, 1950.

[12] E. L. Lehmann. Significance Level and Power. *Ann. Math. Statist.*, Vol. 29, No. 4, pp. 1167-1176, 1958.
[13] E. L. Lehmann. Optimum invariant tests. *Ann. Math. Statist.*, Vol. 30, pp. 881-884, 1959.
[14] E. L. Lehmann. *Testing Statistical Hypotheses*, John Wiley and Sons, New York, pp. xiii + 369, 1959.
[15] E. L. Lehmann. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Amer. Statist. Assoc.*, Vol. 88, pp. 1242-1249, 1993.
[16] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses,* 3^{rd} *Edition*, Springer, 2005.
[17] E. L. Lehmann. On likelihood ratio tests. In *The Second Erich L. Lehmann Symposium Optimality*, (J. Rojo, Ed.), IMS Lecture Notes-Monograph Series Vol. 49, pp. 1-9, 2006.
[18] E. L. Lehmann. *Reminiscences of a Statistician: The Company I Kept*, Springer, 2008.
[19] H. Scheffé. On the Theory of Testing Composite Hypotheses with One Constraint *Ann. Math. Statist.*, Vol. 13, No. 3, pp. 280-293, 1942.